Prédiction de transcriptome : analyse comparative multi-gènes, orthologues

Nicolas Guillaudeux

INRIA (**I**nstitut **N**ational de **R**echeche en **I**nformatique et en **A**utomatique) de
Rennes - Bretagne Atlantique
IRISA (**I**nstitut de **R**echerche en **I**nformatique et **S**ystèmes **A**léatoires)
Équipe Dyliss
*(**DY**namics, **L**ogics and **I**nference for biological **S**ystems and **S**equences)*

campus de Beaulieu
263 avenue Général Leclerc
35042 Rennes

Prédiction de transcriptome : analyse comparative multi-gènes, orthologues

Rapport de stage

Master 2 Bio-informatique
parcours Informatique et Biologie Intégrative
Année universitaire 2017/2018

Encadrante :
Catherine Belleannée

Auteur :
Nicolas Guillaudeux

Partenaires du stage :
Samuel Blanquart
Jean-Stéphane Varré

Master bioinformatique

UNIVERSITÉ DE RENNES 1

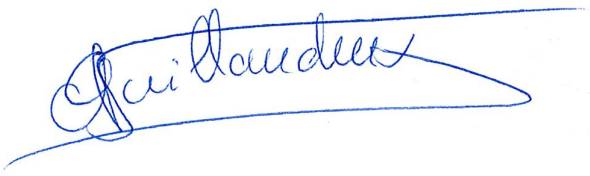
ENGAGEMENT DE NON PLAGIAT

Je, soussigné(e) ..Nicolas..Guillaudeux.....
étudiant(e) en..M2..de..Bioinformatique.....
déclare être pleinement informé que le plagiat de documents ou d'une partie de document publiés sur toute forme de support, y comprit l'internet, constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée.

En conséquence, je m'engage à citer toutes les sources que j'ai utilisées pour la rédaction de ce document.

Date : 13/06/2018

Signature :



Document à compléter de manière manuscrite et à insérer obligatoirement en première page du rapport de stage.

Remerciements

Je tiens à remercier Catherine Belleannée pour m'avoir accordé ce stage, pour son aide et ses conseils, sa disponibilité tant sur le stage que sur l'accompagnement au cours du stage.

Je remercie aussi Samuel Blanquart et Jean-Stéphane Varré pour leur disponibilité et leurs conseils qui m'ont permis de mieux m'approprier l'outil CG-Alcode mais aussi pour leurs conseils sur le développement des méthodes mises en route pour le déroulement de l'étude.

Je remercie de nouveau ces trois personnes pour le temps qu'elles m'ont accordé pour la rédaction et la relecture de ce présent rapport.

Enfin, je remercie l'équipe Dyliss pour son accueil chaleureux mais aussi mes collègues stagiaires à l'IRISA pour leur bonne humeur, le cadre de travail convivial et leur entraide qui mérite un point spécial.

Abréviations

ARN	Acide ribonucléique
ARNm	Acide ribonucléique messager
CCDS	Consensus Coding Sequence
CG-ALCODE	Comparative genomics for alternative coding in eukaryote genes
CLF	*Canis lupus familiaris*
CREM	cAMP Responsive Element Modulator
EST	Expressed Sequence Tag
GA	Gene Actual
GEXF	Graph Exchange XML Format
GO	Gene Oracle
HS	*Homo sapiens*
IN_END	Évènement de fin d'intron (accepteur d'épissage)
IN_START	Évènement de début d'intron (donneur d'épissage)
JSON	JavaScript Object Notation
MM	*Mus musculus*
NGS	Next Generation Sequencing
ORF	Open Reading Frame
RNA-seq	RNA-sequencing
TL_END	Évènement de fin de traduction (codon *stop*)
TL_START	Évènement de début de traduction (codon *start*)
UTR	Untranslated Transcribed Region
YAML	Yet Another Markup Language

Table des matières

1 Introduction

1.1 Contexte de l'étude

1.1.1 Étapes de l'expression des gènes chez les eucaryotes

Chez les organismes eucaryotes, l'information portée par les gènes est exprimée par un mécanisme d'expression divisé en plusieurs étapes hautement régulées. La première étape consiste en la transcription du gène en un produit intermédiaire, le **pré-ARNm**. Celui-ci subit ensuite une étape de maturation déclinée en trois sous-étapes que sont le coiffage, la polyadénylation et l'**épissage**. Ensemble, ces étapes permettent de produire un **ARN messager (ARNm) mature** prêt à être traduit en **protéine**, le produit final qui assure une fonction. On distingue ainsi les étapes de transcription, de maturation et de traduction. La production du pré-ARNm est sous contrôle d'une régulation dite transcriptionnelle et la maturation du pre-ARNm en ARNm est sous contrôle d'une régulation post-transcriptionnelle.

1.1.2 Transcrits alternatifs et isoformes : un gène, plusieurs pré-ARNm, ARNm et protéines

La séquence d'un gène peut contenir plusieurs promoteurs alternatifs régulés de manière à initier chacun la transcription par une extrémité 5'-UTR particulière de pré-ARNm. De même, des sites de polyadénylation alternatifs permettent de terminer chacun la transcription avec une extrémité 3'-UTR particulière. Ce mécanisme est nommé **transcription alternative** [1] (figure 1A).

Au cours de la maturation de ces pré-ARNm, l'étape d'épissage est assurée par un complexe ribonucléoprotéique, le spliceosome. Elle consiste à exciser des segments du pré-ARNm, les **introns**, et à rabouter les autres régions, les **exons**, pour former des ARNm matures ou **transcrits**. Ainsi, les exons correspondent aux régions qui forment l'ARNm mature. Le spliceosome est un complexe qui se met en place au travers de **sites fonctionnels** contenus sur la séquence du pré-ARNm tels que les **sites donneurs et accepteurs d'épissage** ou encore le point de branchement. [2] Les exons se classent en deux catégories distinctes : d'une part, les **exons constitutifs** qui correspondent à des exons toujours présents dans les transcrits matures, et d'autre part les **exons alternatifs** qui ne sont pas toujours sélectionnés par le processus d'épissage. Des étapes de régulation entraînent une sélection sur ces exons alternatifs. On parle ainsi de régulation post-transcriptionnelle ou d'**épissage alternatif** [2–5]

(figure 1B). Cette sélection provoque ainsi la formation de **transcrits alternatifs** ainsi nommés du fait que leur contenu en exons ne sera pas identique. Les protéines traduites à partir des transcrits alternatifs d'un même gène sont appelées **protéines isoformes** (figure 1C).

Ces mécanismes jouent un rôle important dans l'expansion de la diversité protéique et principalement chez les vertébrés [6]. En effet, on estime que 95% des gènes sont concernés par ce processus d'épissage alternatif [7]. Cette sélection d'exons peut-être influencée par des facteurs protéiques [8], des petits ARN non-codants [9] ou encore par le repliement d'un site d'épissage sur un pré-ARNm le rendant inaccessible aux facteurs d'épissage [10]. Ces éléments de régulation de l'épissage alternatif peuvent être mobilisés de différentes manières selon le tissu où s'exprime le gène, ce qui participe à la spécificité des isoformes aux tissus [11].

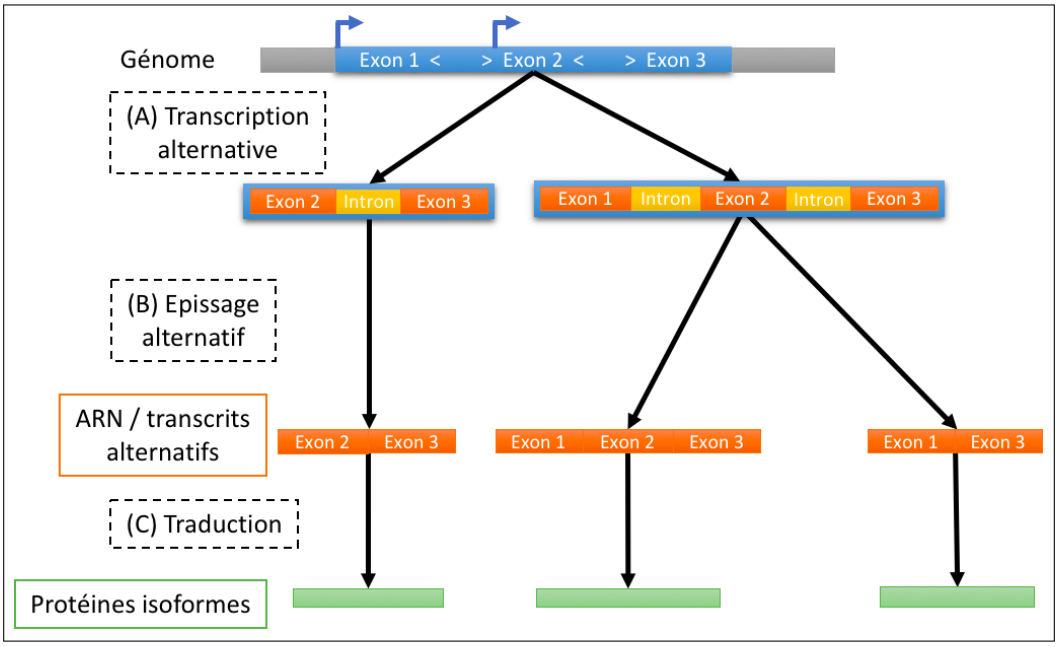


FIGURE 1 – Influence de l'épissage alternatif sur la formation des protéines chez les eucaryotes. Un même gène eucaryote peut posséder plusieurs promoteurs alternatifs (*flèches bleues*) qui par transcription alternative (A) conduisent à la formation de plusieurs pré-ARNm composés de différents exons *(blocs orange)* et introns *(blocs jaune)*. La maturation entraîne une excision des introns et un raboutage des exons pour donner des ARN messagers matures ou transcrits alternatifs, c'est l'épissage (B). Sur cette figure, l'exon 2 est un exon alternatif qui peut être ou non sélectionné permettant la formation de deux types de transcrits alternatifs. Ces ARNm alternatifs donneront des protéines isoformes par traduction (C). Ainsi, une même région du gène peut être transcrite ou non, épissé ou non, traduite ou non.

1.2 Transcriptome d'une espèce

La connaissance de l'ensemble des transcrits qui peuvent être exprimés à partir d'un gène chez une espèce, nommé **transcriptome**, est nécessaire pour identifier l'origine de maladies [5,11] telles que le syndrome de Riley-Day ou dysautonomie familiale [12]. Dans cet exemple, la modification d'un seul nucléotide T en un nucléotide C dans l'intron 20 entraîne la non sélection de l'exon 20 pour le transcrit mature. Ceci conduit à la formation d'un transcrit avec un codon *stop* prématuré qui va contribuer à de graves complications neurologiques. Cet exemple illustre l'utilité d'identifier l'ensemble des transcrits d'un gène. Pour cela, différentes méthodes se sont succédées [13, 14] :

- analyses comparatives basées sur des étiquettes de séquences exprimées (EST) : cette méthode permet d'identifier certains évènements d'épissage mais fournit des données incomplètes et est très coûteuse,
- analyses sur des puces à ADN (*microarray*) : cette méthode permet d'identifier de nombreux évènements d'épissage alternatif mais est limitée par la densité, le *design* de la séquence et l'analyse des données reste difficile [11, 13],
- analyses via les données de séquençage de l'ARN (RNA-seq) : ces méthodes se sont davantage développées à la suite de la diminution des coûts des séquenceurs de nouvelle génération (NGS) et sont majoritairement utilisées aujourd'hui [7].

Ces technologies expérimentales (et d'autres) renseignent nos connaissances sur les transcriptomes, produisant des transcrits, parfois incomplets, que nous appellerons **transcrits connus** par la suite.

1.2.1 Méthode standard : assemblage des transcrits par RNA-seq

Grâce au développement des NGS, de nombreux logiciels ont été développés et utilisent les données issues des RNA-seq [13–17]. Ces méthodes ont pour but de reconstruire les transcrits complets à partir des lectures (*reads*) issues du séquenceur en utilisant des données expérimentales au départ. Ces méthodes doivent se baser sur un génome de référence pour pouvoir être assemblées réduisant ainsi les possibilités d'assemblage *de novo* [7]. De plus, les données RNA-seq introduisent des biais compliquant leur analyse. (1) En effet, les lectures produites sont beaucoup plus courtes que les transcrits complets. Si elles permettent d'identifier les jonctions d'exons consécutifs, elles ne permettent pas facilement d'identifier comment des exons distants (d'une distance supérieure à la taille de la lecture) composent ou non un transcrit complet. (2) Par ailleurs, les lectures courtes obtenues aux extrémités 5' et 3'-UTR du transcrit sont généralement filtrées, ce qui complique l'assemblage de transcrits complets. Enfin, l'infor-

mation de couverture en lectures est généralement utilisée pour inférer un transcrit complet, la couverture devant théoriquement rester homogène tout au long de l'assemblage. Or, dans les données expérimentales cette couverture varie (toutes les zones du transcrit ne sont pas amplifiées de manière homogène, ce qui biaise l'estimation de transcrits complets faiblement exprimés). Une étude de 2013 [18] a permis de montrer que la meilleure des méthodes d'assemblage de transcrits complets basées sur les données de RNA-seq n'était capable de reconstruire que 21% des transcrits humains connus.

1.2.2 CG-Alcode : prédiction de transcrits par conservation structurelle entre deux espèces

Du fait des limites des méthodes développées jusqu'à présent, d'autres approches complémentaires sont nécessaires pour mettre en évidence l'ensemble de ces transcrits. Le développement d'une méthode visant à rechercher les délimitations intron/exon en se basant sur la séquence du gène semblerait être une nouvelle approche. Seulement, l'utilisation des délimiteurs d'épissage qui sont de courts motifs dinucléotidiques (AG et GT) n'est pas suffisante pour repérer les zones introniques. En effet, ces motifs sont abondants au sein des séquences et ne sont pas discriminants. Ainsi, une recherche *ex nihilo* de ces motifs entrainerait une explosion combinatoire. Pour dépasser cette limite, une méthode de prédiction des transcrits, assistée par des connaissances biologiques, a été mis en place à Lille par les partenaires de ce stage sous le nom de CG-Alcode [19].

La méthode CG-Alcode exploite le principe de conservation de séquences d'un même gène entre deux espèces (dit "gènes orthologues"). On définit deux **gènes orthologues** de deux espèces comme étant deux copies héritées d'un même gène qui était présent chez leur ancêtre commun. CG-Alcode considère ainsi le gène g1 d'une espèce E1, dit **gène source**, et son orthologue g2 sur une espèce E2, dit **gène cible**, et tente de transposer les informations connues de g1 sur g2. En particulier, CG-Alcode étudie si les transcrits connus de g1 pourraient être produits par g2. Si c'est le cas, ces **transcrits** sont **prédits** chez g2. La méthode repose sur l'alignement des exons connus sur le gène source avec le gène cible afin de compléter les informations connues sur le gène cible.

1.3 Objectifs du stage

L'objectif de cette étude est de proposer une méthodologie pour comparer les transcriptomes d'un gène sur plusieurs espèces et d'en extraire de l'information phylogénétique. Il s'agit donc de compléter la méthode CG-Alcode actuelle au-delà de la comparaison entre l'homme et

la souris. L'idée est d'augmenter le nombre de transcrits prédits et de permettre des analyses phylogénétiques. L'étude portera sur l'homme, la souris et le chien. Le travail se décompose en deux étapes principales :

1. Compléter les connaissances disponibles pour une paire de gènes (*pairwise*) entre deux espèces afin d'affiner la connaissance de la structure du gène et les transcrits prédits. CG-Alcode peut annoter un génome non modèle. La première étape nous permettra notamment d'augmenter la prédiction sur le génome non modèle du chien en tirant profit des connaissances disponibles sur les transcriptomes modèles humain et murin.
2. Mettre en place une comparaison multi-espèces pour permettre des analyses phylogénétiques. Le chien est une espèce extérieure au groupe de l'homme et de la souris, les données et prédictions issues du génome canin permettent de mieux comprendre l'évolution des gènes humains et murins. CG-Alcode fonctionne pour une paire de gènes source / cible, l'approche ne permet pas actuellement de comparer un même gène sur plusieurs espèces. Nous allons donc développer une méthode pour permettre une comparaison multi-espèces des gènes orthologues considérés chez l'humain, la souris et le chien.

2 Matériel et Méthodes

2.1 Données utilisées : transcrits exprimés par l'homme, la souris et le chien

Dans leur article [19], Samuel BLANQUART et Jean-Stéphane VARRÉ, partenaires de ce stage, ont analysé une cohorte de 1 936 paires de gènes chez l'humain et la souris exprimant 15 513 transcrits alternatifs connus. Les résultats détaillés de la méthode ont été illustrés sur le cas du gène CREM, un gène exprimant des facteurs de transcription régulant la transcription de plusieurs gènes. CG-ALCODE a été testé sur un total de 1 936 paires de gènes et la meilleure cohorte considérée atteint 91% de transcrits correctement prédits.

Les gènes et les transcrits de trois espèces sont utilisés dans le cadre de cette étude. Parmi elles, deux sont des espèces modèles dont les données sont référencées dans la base de données CCDS [20] (version 21 du 12/2017) : l'homme (*Homo sapiens*, hs) et la souris (*Mus musculus*, mm). La troisième espèce est une espèce non modèle : le chien (*Canis lupus familiaris*, clf) dont les données proviennent directement d'*Ensembl* [21] (version 91 du 12/2017).

Gènes et transcrits : Seuls les gènes présents en copie unique sur les trois espèces (ayant un lien orthologique "*1to1*") et ayant au moins deux transcrits alternatifs dans CCDS ont été conservés. Au total, **2 167 triplets de gènes** et **18 151 transcrits connus** composent le jeu de données de cette étude.

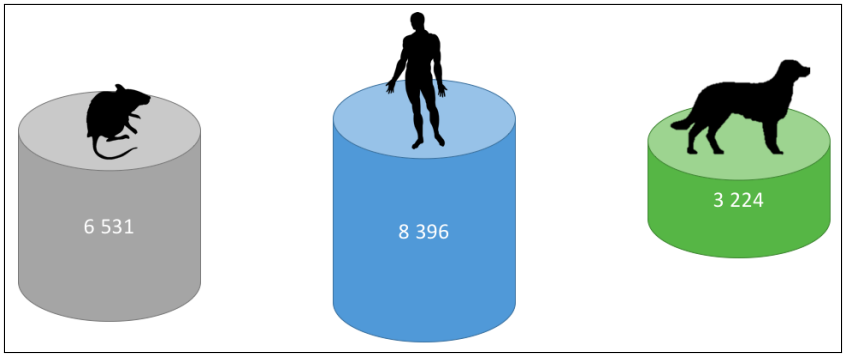


FIGURE 2 – Répartition des transcrits connus entre les espèces. 2 167 gènes orthologues expriment 8 396 transcrits connus chez l'homme, 6 531 transcrits connus chez la souris et 3 224 chez le chien.

Gène CREM : Au cours de ce stage, le gène CREM a de nouveau été utilisé pour tester les implémentations ajoutées à la méthode CG-ALCODE et pour en illustrer les résultats détaillés,

et notamment d'un point de vue phylogénétique. **Le gène CREM exprime 21 transcrits connus chez l'homme, 13 chez la souris et 3 chez le chien**.

2.2 Logiciels utilisés

2.2.1 CG-ALCODE

Ce stage vise à poursuivre le développement de la méthode CG-ALCODE [19]. CG-ALCODE est un logiciel qui utilise les transcrits connus et référencés d'un gène chez une espèce **source** pour tenter de les transposer à un **gène cible** orthologue d'une espèce voisine de manière à pouvoir prédire les transcrits alternatifs et les protéines isoformes que le gène cible est capable d'exprimer (ou d'"exécuter"), on parle alors de **transcrits prédits**. Ainsi, CG-ALCODE est définie comme une méthode de *reconstruction assistée du transcriptome*.

Données en entrée : Le programme utilise deux ensembles de données en entrée : le gène et ses transcrits connus chez l'espèce source et le gène orthologue et ses transcrits connus chez l'espèce cible.

Données en sortie : A partir de ces données, le programme crée une grammaire spécifique sous forme de blocs et de sites fonctionnels pour définir un modèle du gène commun aux deux espèces et pour définir une notation des transcrits connus et prédits [22]. Ainsi, un bloc correspond à une portion génomique délimitée par deux sites fonctionnels et est représenté par une lettre lorsqu'il s'agit d'un exon ou par un point "." lorsqu'il s'agit d'un intron. Les codons *start* et *stop* sont représentés respectivement par les symboles "[" et "]". Les sites donneurs (GT) et accepteurs (AG) d'épissage, qui délimitent les introns, sont représentés respectivement par les symboles "<" et ">" (figure 3, tableau 1).

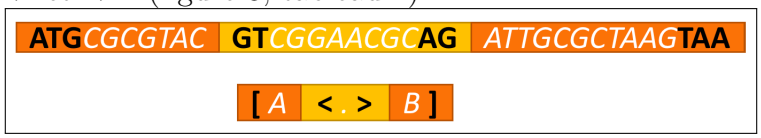


FIGURE 3 – Représentation abstraite d'un gène sous CG-ALCODE. Dans cet exemple, le codon *start* (ATG) et le site donneur d'épissage (GT) sont des sites qui délimitent un bloc exonique nommé "A". Le site donneur (GT) et le site accepteur (AG) d'épissage délimitent un bloc intronique noté ".". Le site accepteur d'épissage (AG) et le codon *stop* (TAA) délimitent un bloc exonique nommé "B".

2.2.1.a Définition des modèles GA et GO

A partir du répertoire des transcrits connus d'un gène, le programme les projette sur le gène pour délimiter les régions intron/exon et définir des blocs (figure 4 étape 1). Cette étape est réalisée à la fois pour le gène source et pour le gène cible de façon indépendante. L'ensemble des blocs et sites fonctionnels définit un modèle des éléments connus du gène : *gene actual (GA)*. Le modèle GA contient ainsi l'ensemble des informations qui ont été apportées par

Tableau 1 – Réprésentation symbolique et nommage des sites fonctionnels. Quatre types de sites fonctionnels sont considérés dans la méthode CG-Alcode.

Représentation	Nom	Type de site fonctionnel	Valeur
[	TL_START	codon *start*	ATG
]	TL_END	codon *stop*	TAA, TAG, TGA
>	IN_START	donneur d'épissage	GT
<	IN_END	accepteur d'épissage	AG

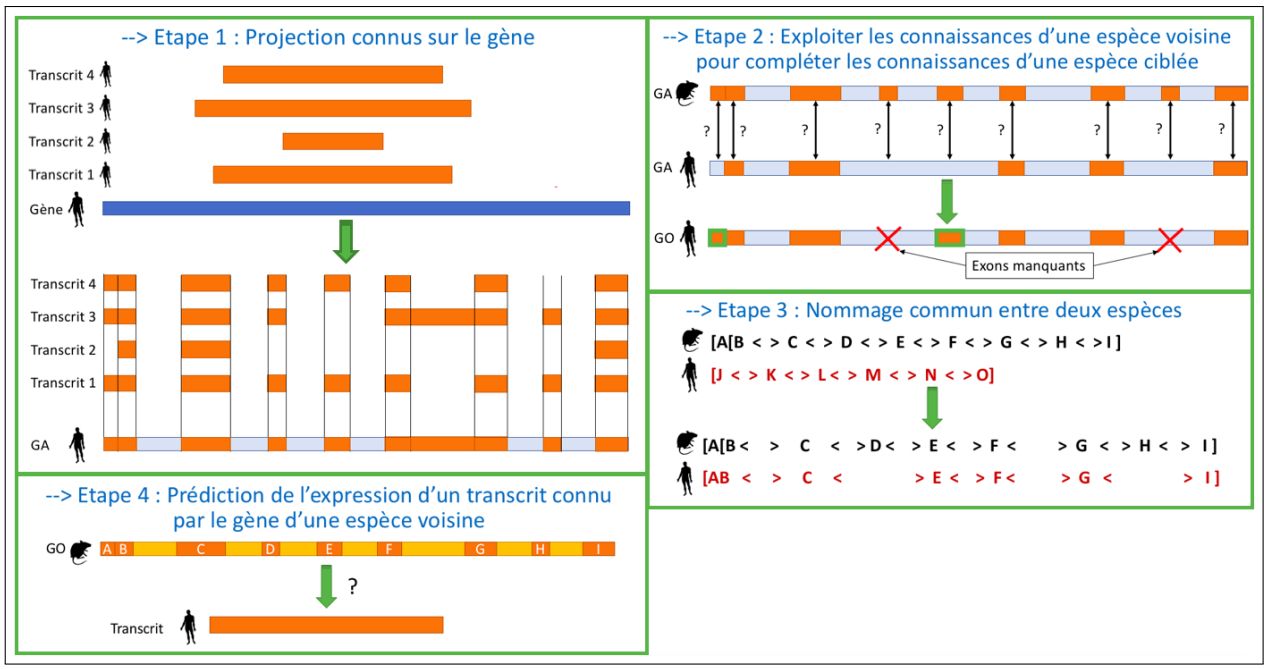


Figure 4 – Pipeline de la méthode CG-Alcode pour la définition des modèles GA et GO chez deux espèces. Étape 1. Projection des transcrits connus d'une espèce sur son gène pour délimiter les blocs introniques et les blocs exoniques définissant le modèle *Gene Actual* (étape réalisée indépendamment pour les deux espèces). **Étape 2**. Alignements de chacun des blocs du gène de l'espèce source sur le gène entier de l'espèce cible (et réciproquement) à la recherche d'une homologie entre les séquences. Les blocs qui ne sont pas alignés (croix rouge) sont spécifiques à l'espèce source : l'espèce cible ne contient pas ces exons. Cet ajout d'information fait apparaître de nouveaux blocs sur le gène cible et définit le modèle *Gene Oracle*. **Étape 3**. Les noms des blocs sont mis en commun entre les modèles. Les blocs orthologues sont nommés de la même façon et les blocs spécifiques ont un nommage qui leur est propre. De plus, les signaux délimitant ces blocs sont également estimés orthologues. **Étape 4**. Vérification de la modélisation et prédiction de l'expression d'un transcrit source par le gène cible.

les transcrits du gène. La figure 5 représente le résultat (après nommage commun) de cette projection chez l'homme et la souris pour le gène CREM.

Chaque bloc de l'espèce source est ensuite aligné par Blast [23] contre toute la séquence nucléotidique du gène de l'espèce cible en utilisant un contexte de vingt nucléotides additionnels

```
GA homme  [A<.>B[C<.[D<.[E<.    >G<.>H<IJ].>K[L<. [N<.[O<.  >Q[R<.>S<.>T].>UV]
GA souris [A<.>B[C<.    [E<.>F<.>G<.>H<I] .>K L<.[MN<.    [P>Q[R<.>S<.>T].>U]
```

Figure 5 – Modèles *gene actual* du gène CREM chez l'homme et la souris. L'ensemble des blocs et sites fonctionnels définit un modèle des éléments connus du gène, sur l'espèce indiquée. Ainsi un bloc X présent sur GA homme indique qu'au moins un des transcrit connus humains contient X.

```
GO homme  [A<.>B[C<.[D<.[E<.>F<.>G<.>H<IJ].>K[L<. M[N<.[O<.P>Q[R<.>S<.>T].>U V]
GO souris [A<.>B[C<.[D<.[E<.>F<.>G<.>H<I]J.>K[L<.[M[N<.   [P>Q[R<.>S<.>T].>U]V]
```

Figure 6 – Modèles *gene oracle* du gène CREM chez l'homme et la souris. La composition de ces modèles révèle la capacité du gène d'une espèce à exprimer une prédiction de présence / absence des transcrits. Ainsi un bloc X présent sur GA homme indique que soit un des transcrit connus humains contient X ou soit qu'un des transcrits murins contient X et que celui-ci a une relation d'orthologie dans le gène humain.

de part et d'autre du bloc de manière à aligner les parties conservées des sites d'épissage (figure 4 étape 2). Le score de E-value pour les Blast est calibré à 10^{-4}. Ces alignements permettent ainsi de voir si une homologie existe pour les blocs et les sites fonctionnels entre les deux espèces. Si une homologie est retrouvée, on aura donc une **relation d'orthologie** entre les deux gènes orthologues pour le bloc aligné et les sites fonctionnels qui le flanquent. Il est important de noter que deux blocs connus (*actual*) peuvent être alignés ensemble. L'orthologie appliquée aux blocs et aux sites fonctionnels indique qu'ils étaient présent chez l'ancêtre commun et qu'ils sont partagés par les gènes orthologues des espèces filles. Ainsi, les blocs qui n'étaient présents que chez l'une des deux espèces et qui ont une relation d'orthologie dans le gène de l'autre espèce sont prédits. Ces blocs prédits complètent le modèle GA qui est alors qualifié de modèle oracle du gène (*gene oracle (GO)*) capable de produire une prédiction de présence / absence des transcrits. Dans la figure 5, les blocs D, J et V humains absents des modèles GA de la souris sont retrouvés présents et figurent dans le modèle GO de la souris (figure 6). Ce cas est aussi réciproque pour les blocs F, M et P qui ne sont pas dans le modèle GA de l'homme mais qui ont une relation d'orthologie prédite d'où leur présence dans le modèle GO de l'homme. Le bloc O humain qui n'a pas d'orthologie prédite chez la souris, d'où son absence dans le modèle GO de la souris.

Une fois les modèles GO définis, le programme nomme chacun des blocs de façon commune entre les deux espèces. Deux blocs orthologues auront ainsi la même dénomination alors qu'un bloc spécifique à une espèce aura une dénomination propre (figure 4 étape 3). Ces modèles sont considérés comme valides si il n'y a pas de réarrangement de l'ordre des blocs (colinéarité des blocs orthologues) et qu'il n'y a pas de blocs dupliqués à l'identique.

Ainsi chaque bloc / site dans un modèle de gène peut être associé avec un autre bloc / site orthologue dans l'autre gène. Cette association définit une **relation d'orthologie** entre blocs ou entre sites dans les modèles des gènes chez les deux espèces. Ces relations obtenues pour une paire de gènes seront exploitées ultérieurement pour réaliser la comparaison multiple des gènes chez l'homme, la souris et le chien.

2.2.1.b Exécutabilité des transcrits

A partir du modèle GO d'une espèce cible et des transcrits connus chez l'espèce source, le logiciel prédit si ces transcrits peuvent être exprimés chez l'espèce cible. Si tous les blocs et les sites d'épissage nécessaires à l'expression d'un transcrit source dans un gène source sont présents dans le modèle GO du gène cible alors le transcrit source est dit exécutable d'un point de vue grammatical par le gène cible. Autrement dit, le programme prédit que tous les sites fonctionnels requis étant présents dans le gène cible, alors un transcrit orthologue du transcrit source peut-être exprimé par le gène cible (figure 4 étape 4 et figure 7).

A partir du moment où le transcrit source est exécutable par le gène cible, le programme cherche la présence du transcrit orthologue prédit dans les transcrits connus de l'espèce cible. Dans un premier temps, il recherche si une structure grammaticale identique du transcrit source existe dans les transcrits connus du gène cible. Si une correspondance existe, le transcrit prédit est dit "*found*" (double flèche sur la figure 7) : on a trouvé un transcrit connu correspondant au transcrit prédit. Si le transcrit source n'a pas de correspondance dans les transcrits cibles alors il est qualifié de "*yet-to-be-found*" (simple flèche sur la figure 7) : on estime que le transcrit prédit sera observé dans les produits d'expression du gène cible. En plus de réaliser une vérification grammaticale, le programme cherche à savoir si ces transcrits prédits sont réellement traductibles dans leur cadre de lecture (ORF ou région codante). Si le cadre de lecture est valide (débute par un codon *start*, se termine par un codon *stop* et sa longueur est un multiple de 3), le transcrit est qualifié comme "*Correct-ORF*". A l'inverse, si un transcrit n'est pas retrouvé exécutable d'un point de vue syntaxique et/ou si son ORF est incorrect, celui-ci est dit non exécutable ("*No-executable*").

Ainsi cette relation d'exécutabilité des transcrits source et cible amène à la notion de **relation d'orthologie entre transcrits**. Si la relation d'orthologie entre transcrits est "*found*", alors le transcrit source connu possède un transcrit orthologue connu dans le gène cible (et réciproquement). Si la relation d'orthologie entre transcrits est "*yet-to-be-found*" alors le transcrit source connu possède un transcrit orthologue inconnu prédit dans le gène cible.

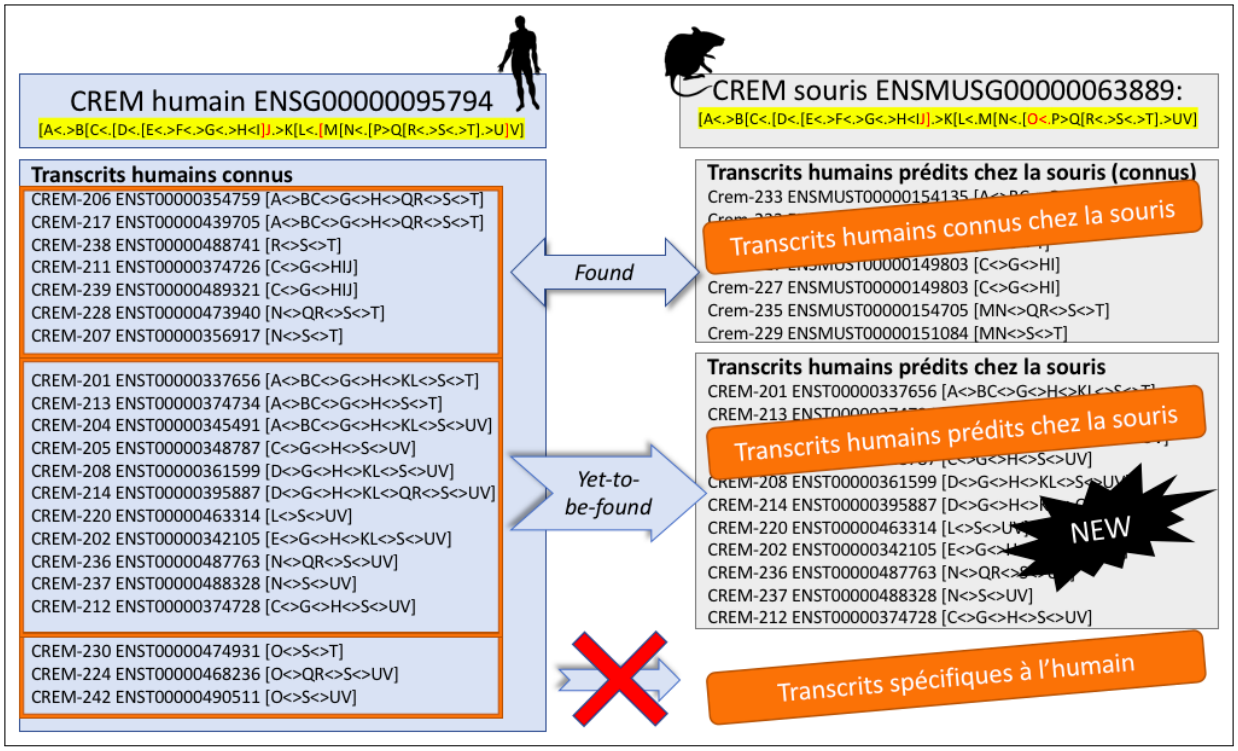


Figure 7 – Exécutabilité des transcrits connus entre l'homme et la souris. 21 transcrits alternatifs sont connus pour le gène CREM humain, dont 18 sont prédits exprimables (exécutables) par le gène CREM de la souris. Parmi eux 7 ont un transcrit orthologue connu chez la souris et 11 sont des transcrits prédits, qui pourraient être exprimés chez la souris. 3 transcrits humains ne sont pas exprimables par le gène CREM murin et sont spécifiques au gène CREM humain.

2.2.2 NetworkX : création et manipulation de graphes

NetworkX [24] est un package Python utilisé pour générer, manipuler et étudier des graphes. Il permet de générer des graphes dirigés ou non et possède des méthodes d'exportation de graphes dans des formats de sortie tels que *YAML*, *JSON* ou encore *GEXF*. Dans le cadre de cette étude, le format *GEXF* a été choisi pour les fichiers de sortie notamment dans le cas de la visualisation avec le logiciel Gephi.

Les graphes de sites fonctionnels orthologues chez les trois espèces (homme, souris et chien) générés au cours de ce stage ont été faits de la façon suivante (figure 9) : les nœuds représentent des positions génomiques et le type de site fonctionnel. Lorsqu'un site fonctionnel d'une espèce source possède une relation d'orthologie chez une espèce cible, alors un lien dirigé est créé du nœud de l'espèce source vers le nœud de l'espèce cible. La dénomination des nœuds est donnée dans le tableau 1.

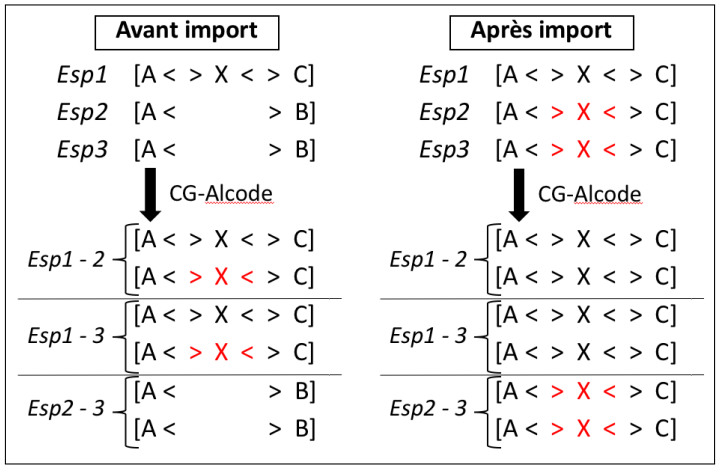


Figure 8 – Méthode d'import des transcrits prédits. La partie gauche représente le résultat des comparaisons de paires de gènes réalisées avant import des prédictions. Les relations d'orthologie entre les trois gènes ne peuvent pas être comparées directement : l'orthologie de X dans les gènes des espèces Esp2 et Esp3 n'est pas estimée lors de la comparaison Esp2 - 3 et le graphe de relations d'orthologie ne pourra pas indiquer que ces gènes partagent aussi l'exon X. La partie droite représente le résultat des comparaisons de paires de gènes après import des prédictions. Ainsi la comparaison Esp2 - 3 prend en compte une information nouvelle issue transitivement de la comparaison de ces gènes avec le gène de l'espèce Esp1.

2.2.3 Gephi : visualisation de graphes

Gephi [25] est un logiciel de visualisation et d'exploration de graphes. Il a été utilisé pour réaliser des rendus visuels des graphes générés au cours de cette étude. Les paramètres de l'outil utilisés concerne l'algorithme de spatialisation Force Atlas avec une répulsion de 1000 et une attraction de 0.05 suivi d'un ajustement des labels.

2.3 Contribution : méthode d'export des transcrits prédits

Pour enrichir les comparaisons par paire d'espèces et obtenir ainsi des modèles de gènes plus précis, nous avons mis en place une méthode d'export des transcrits prédits. Habituellement pour comparer deux gènes orthologues sur deux espèces E1 et E2, CG-Alcode s'appuie sur les transcrits connus sur chacune des espèces. L'idée ici est d'ajouter aux transcrits de départ connus chez E1 et E2 les transcrits prédits à partir d'une autre espèce E3. Ceci permettra de faire apparaître des blocs et des signaux connus spécifiquement chez E3 dans les gènes de E1 et E2. La figure 8 montre un exemple où l'export de transcrits prédits provenant de E1 enrichit la comparaison entre E2 et E3 et y fait apparaître un exon, X, qui provient spécifiquement des connaissances issues de E1.

De nouvelles méthodes ont été implémentées dans CG-Alcode pour exporter les transcrits prédits dans les répertoires des espèces. La principale fonction implémentée est la fonction

exportPredictionAsENSEMBL. Celle-ci prend en entrée le transcrit prédit "*yet-to-be-found*" ayant un ORF valide et le gène du transcrit prédit. La fonction exporte le transcrit prédit c'est à dire qu'un fichier transcrit est écrit.

Lorsqu'un transcrit prédit peut-être obtenu à partir de deux transcrits sources distincts, une seule source est arbitrairement choisie et le transcrit prédit n'est exporté qu'une seule fois.

L'ensemble de ces transcrits exportés par le programme est ensuite réinjecté dans les jeux de données des espèces.

2.4 Contribution : graphes d'orthologie entre sites fonctionnels

Le second axe d'étude s'oriente vers de la comparaison multi-gènes et multi-espèces dans la finalité d'obtenir le même nommage dans les modèles de gènes. Dans ce but, les graphes d'orthologie entre sites fonctionnels sont construits pour montrer les relations d'orthologie estimées entre les sites fonctionnels d'un triplet de gènes orthologues. Ces graphes sont construits à partir des modèles GO définis lors des comparaisons des paires de gènes après réinjection des prédictions. Les graphes générés par NetworkX sont exportés au format "*.gexf*" pour pouvoir être visualisé sous le logiciel Gephi.

Si un site fonctionnel est partagé de façon réciproque entre les trois gènes des trois espèces, une clique à trois sommets est tracées (***clique3***). Si une information est partagée réciproquement seulement entre deux gènes de deux espèces, c'est une clique à deux sommets qui est tracée (**doublon**). Enfin, une information spécifique au gène d'une espèce sera représentée par un **singleton** (figure 9).

Ainsi, une collection de sous-graphes représentera l'information relative aux structures d'un triplet de gènes orthologues, chaque sous-graphe indiquant de quelle manière un site fonctionnel est partagé par les gènes.

Si les graphes de sites fonctionnels représentant les relations d'orthologie de tous les sites fonctionnels d'un triplet de gènes sont entièrement composés de *clique3*, alors tous les sites fonctionnels sont partagés par les trois gènes et par conséquent ils étaient vraisemblablement déjà présents chez l'ancêtre. De plus, dans ce cas chaque comparaison de paires de gènes donne des modèles GO identiques et on peut en déduire que le modèle GO obtenu est le modèle commun aux trois gènes.

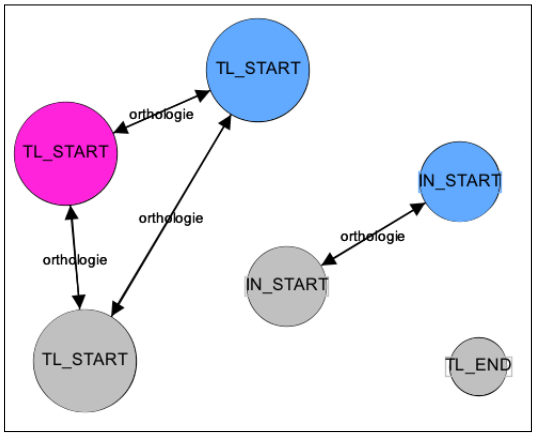


FIGURE 9 – Représentation utilisée pour la création des graphes de relations d'orthologie des sites fonctionnels. Chaque nœud correspond à un site fonctionnel porté par un gène. Les flèches indiquent les relations d'orthologie entre les sites fonctionnels des gènes des trois espèces. Les nœuds chez l'homme sont colorés en bleu, les nœuds chez la souris en gris et les nœuds chez le chien en rose. Trois sous-graphes correspondant chacun à un locus distinct des gènes orthologues sont montrés, dont les topologies sont *clique3*, doublon ou singleton. Une *clique3* induit donc que ce locus était présent chez l'ancêtre commun de ces trois espèces. Le doublon induit ici que ce locus est apparu chez l'ancêtre *Euarchontoglires* ou bien chez l'ancêtre *Eutheria* et perdu par le chien. Le singleton induit ici un locus spécifique uniquement retrouvé chez la souris. Cette figure indique 3 loci distincts chez la souris, 2 chez l'humain et 1 chez le chien.

2.4.1 Scénarios phylogénétiques

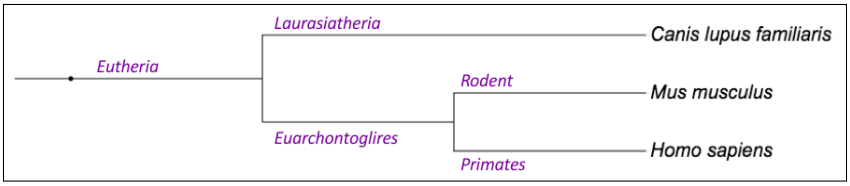


FIGURE 10 – Relation phylogénétique entre les espèces étudiées. L'homme (*Homo sapiens*) est de plus proche parenté avec la souris (*Mus musculus*) qu'avec le chien (*Canis lupus familiaris)*. L'ancêtre commun homme-souris est *Euarchontoglires* et l'ancêtre commun aux trois espèces est *Eutheria*. Si un site fonctionnel est partagé par les trois espèces, alors ce site fonctionnel existait sans doute chez l'ancêtre commun *Eutheria*.

En fonction des graphes obtenus, seuls les graphes n'étant composés que de sous-graphes de type *clique3*, doublons et/ou singletons sont conservés. En effet, les autres cas de relations sont ambigus et mélangent des cas de faux-positifs et des cas non explicitement modélisés (par exemple la duplication). Selon la topologie d'un sous-graphe, sept scénarios sont à considérer en tenant compte de l'hypothèse de parcimonie :

— si le site est présent chez les trois espèces (sous-graphe *clique3*), le site est apparu chez l'ancêtre *Eutheria* ou plus ancien (figure 10),

— si le site est absent chez une espèce (sous-graphe doublon) :
  — si absent chez l'humain ou la souris : il s'agit d'une perte, le site est apparu chez l'ancêtre *Eutheria* ou plus ancien (figure 11a-b),
  — si absent chez le chien : soit il s'agit d'une perte chez le chien et le site est apparu chez l'ancêtre *Eutheria* ou plus ancien (figure 11c), soit il s'agit d'une apparition chez l'ancêtre humain-souris (*Euarchontoglires*) ,
— si le site est présent chez une seule espèce (sous-graphe singleton) :
  — si présent que chez l'humain : le site est apparu chez l'ancêtre *Primate* ou plus récemment (figure 11d),
  — si présent que chez la souris : le site est apparu chez l'ancêtre *Rodent* ou plus récemment (figure 11e),
  — si présent que chez le chien : soit le site est apparu chez l'ancêtre *Laurasiatheria* ou plus récemment (figure 11f), soit le site est apparu chez l'ancêtre *Eutheria* ou plus ancien et a été perdu par l'ancêtre *Euarchontoglires*.

Les triplets de gènes sont ainsi classés de la façon suivante : (1) si seulement des *clique3*, la structure de tout le gène est partagée depuis l'ancêtre *Eutheria* ; (2) si ou moins un doublon, le gène a connu au moins une perte (chez l'homme ou la souris) ou une apparition chez l'ancêtre *Euarchontoglires* ; (3) si au moins un singleton, le gène a connu au moins une apparition récente (homme, souris ou chien), ou encore une perte chez l'ancêtre *Euarchontoglires*. Parmi les gènes, comparer le nombre de cas (1) (très forte conservation) et de cas (3) (acquisitions récentes) renseigne sur l'importance de la conservation ou de la création d'exon / transcrits durant l'évolution des trois mammifères considérés.

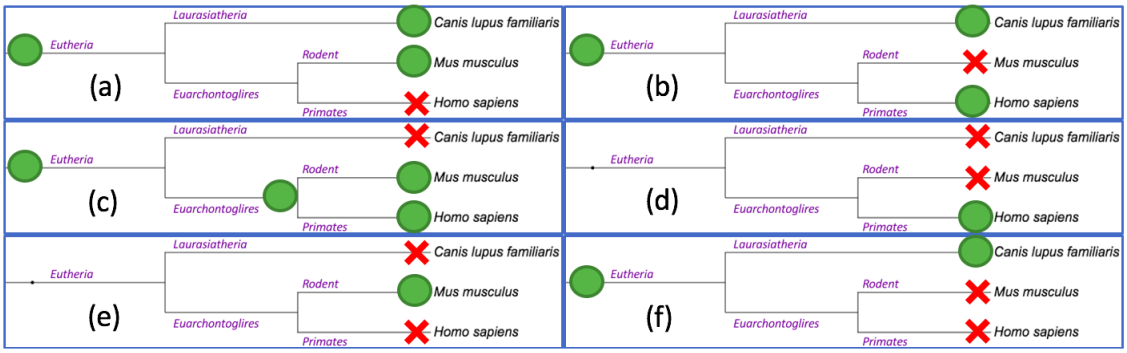


FIGURE 11 – Interprétation phylogénétique des graphes de relations d'orthologie. (a)-(c) Cas des doublons dans les graphes. (d)-(f) Cas des singletons dans les graphes. Un rond correspond à la présence d'un site fonctionnel chez une espèce ou un ancêtre. Une crois correspond à la perte d'un site fonctionnel chez une espèce ou un ancêtre.

3 Résultats et discussion

Nous avons donc développé une méthode pour comparer un même gène chez plusieurs espèces. La première étape est la **réinjection de transcrits prédits**. Elle permet d'enrichir les modèles de paires de gènes grâce à l'import des transcrits prédits à partir de transcrits connus chez d'autres espèces. La deuxième étape est la **comparaison multi-espèces grâce aux graphes d'orthologie de sites fonctionnels**. Nous avons appliqué cette méthode aux données de l'homme, la souris et du chien. Pour chaque étape, nous présentons les résultats en deux temps : (1) d'abord sur le gène d'étude, CREM, (2) puis sur l'ensemble des données.

3.1 Comparaison du gène CREM chez les trois espèces

3.1.1 CREM : trois modèles paire-à-paire

Dans cette étude, CG-Alcode a été appliquée aux données du chien pour ajouter les comparaisons homme-chien (hs-clf) et souris-chien (mm-clf) aux comparaisons homme-souris (hs-mm) initiales [19]. Le tableau 2 indique le résultat de ces comparaisons pour CREM. Pour rappel, le modèle GA est construit à partir des transcrits connus de l'espèce et le modèle GO est complété grâce aux transcrits de la deuxième espèce. Ainsi la comparaison hs-mm permet de faire apparaître 3 nouveaux blocs chez l'homme (F, M, P) et 3 chez la souris (D, J, V). La comparaison hs-clf fait apparaître 4 blocs chez l'homme (B, C, H, T) et 5 chez le chien (F, G, K, P, W). Enfin, la comparaison mm-clf ressort 3 blocs chez la souris (B, C, R) et 5 chez le chien (F, J, N, O, U).

Comme attendu, c'est chez le chien, l'espèce non modèle et la moins documentée, que la comparaison révèle le plus de nouveautés. De plus, chaque espèce est susceptible de prédire de nouveaux blocs chez chaque autre espèce.

Les modèles GO révèlent, de plus, des spécificités. Ainsi, le gène humain possède un bloc non présent chez la souris (O) et un bloc non présent chez le chien (Q). Le gène du chien possède deux blocs non présents chez l'humain (N, O) et deux blocs non présents chez la souris (L, M).

Ces comparaisons de paires de gènes sont indépendantes et les nommages ne sont pas unifiés. Ainsi, un bloc X dans la comparaison hs-mm ne correspondra pas forcément au bloc X dans la comparaison hs-clf ou mm-clf. Il n'est donc pas possible de comparer directement les trois gènes à partir des modèles.

Tableau 2 – Modèles GA et GO obtenus par comparaison de paires de gènes pour l'homme, la souris et le chien. Chaque comparaison de paire de gènes retourne un modèle GA et un modèle GO propre à l'espèce. En rouge figure les blocs exoniques spécifiques d'une espèce au sein d'une comparaison de paire de gènes pour les modèles GO et en bleu les blocs exoniques obtenus grâce aux transcrits de la deuxième espèce.

homme - souris	
GA homme	[A<.>B[C<.[D<.[E<.>G<.>H<IJ].>K[L<.[N<.[O<.>Q[R<.>S<.>T].>UV]
GA souris	[A<.>B[C<.[E<.>F<.>G<.>H<I].>KL<.[MN<.[P>Q[R<.>S<.>T].>U]
GO homme	[A<.>B[C<.[D<.[E<.>F<.>G<.>H<IJ].>K[L<.M[N<.[O<.P>Q[R<.>S<.>T].>UV]
GO souris	[A<.>B[C<.[D<.[E<.>F<.>G<.>H<I]J.>K[L<.[M[N<.[P>Q[R<.>S<.>T].>U]V]
homme - chien	
GA homme	[A<.>D[E<.[F<.[G<.>I<.>J<K].>L[M<.[P<.[Q<.>R[S<.>U<.>V].>W]
GA chien	[A<.[B<.>C>DE<.>H<.>I<.>J<.>LM<.>N<.>O<.>RS<.>T>U<.>V]
GO homme	[A<.B<.>C>D[E<.[F<.[G<.>H<.>I<.>J<K].>L[M<.[P<.[Q<.>R[S<.T>U<.>V].>W]
GO chien	[A<.[B<.>C>D[E<.[F<.[G<.>H<.>I<.>J<K].>L[M<.>N<.>O<.[P<.>R[S<.>T>U<.>V].>W]
souris - chien	
GA souris	[A<.>D[E<.[F<.>G<.>H<.>I<J].>K<.[N<.[O>P[Q<.>S<.>T].>U]
GA chien	[A<.[B<.>C>DE<.>G<.>H<.>I<.>K<.>L<.>M<.>PQ<.>R>S<.>T]
GO souris	[A<.B<.>C>D[E<.[F<.>G<.>H<.>I<J].>K<.[N<.[O>P[Q<.>R>S<.>T].>U]
GO chien	[A<.[B<.>C>D[E<.[F<.>G<.>H<.>I<J.>K<.>L<.>M<.[N<.O>P[Q<.>R>S<.>T].>U]

3.1.2 CREM : transcrits connus et prédits

La comparaison des transcrits connus dans les comparaisons de paires de gènes CREM est présentée dans le tableau 3. La comparaison hs-mm montre que 18 transcrits humains sont exprimables par le gène CREM de la souris dont 7 transcrits humains ont un orthologue connu chez la souris. Quant aux transcrits murins, tous sont exprimables chez l'humain dont 6 ont un orthologue connu chez l'homme.

Parmi les transcrits orthologues connus humain-souris, on note que deux transcrits humains expriment le même ORF (figure 7 de la partie 2.2.1.b)). On a ici un cas de transcription alternative chez l'humain.

Dans la comparaison mm-clf, un transcrit murin a un orthologue connu chez le chien alors que dans la comparaison hs-clf, aucun transcrit connu n'a d'orthologue chez l'autre espèce.

Parmi l'ensemble des comparaisons, 18 transcrits humains connus et 12 chez la souris sont exprimables chez le chien et n'ont pas d'orthologues connus chez le chien. Ces prédictions annotent le transcriptome du chien de façon importante en comparaison avec les prédictions faites sur les deux autres espèces (12 transcrits prédits sont exprimables chez la souris et 9 chez l'humain sans orthologues connus).

Enfin, certains transcrits ne sont pas exprimables chez les deux autres espèces et révèlent des spécificités liées à une espèce comme c'est le cas pour 3 transcrits humains et 1 transcrit canin.

Certains de ces transcrits prédits sont redondants, c'est le cas si le gène humain à un transcrit prédit "*found*" avec un transcrit connu du gène souris et que ces deux transcrits ne sont pas connus chez le chien et qu'ils sont prédits. Pour éviter d'exporter deux fois la même information, ces cas sont filtrés supprimant ainsi 2 doublons chez l'homme, 1 chez la souris et 10 chez le chien. Le tableau 4 présente l'ensemble des transcrits du répertoire de chaque espèce au cours de l'étude. Ainsi, on ajoute 7 transcrits prédits chez l'homme, 11 chez la souris et 20 chez le chien.

Tableau 3 – Transcrits prédits selon les comparaisons de paires de gènes. Ce tableau représente les résultats de projection des transcrits de l'espèce source sur l'espèce cible. Pour la comparaison espèce 1 -> espèce 2, les chiffres indiquent les résultats d'expression des transcrits chez l'espèce 2.

	hs -> mm	**mm -> hs**	**hs -> clf**	**clf -> hs**	**mm -> clf**	**clf -> mm**
Found	7	6	-	-	1	1
Yet-to-be-found	11	7	18	2	12	1
No-executable	3	-	3	1	-	1
Nombre total de transcrits prédits	**11**	**7**	**18**	**2**	**12**	**1**

Tableau 4 – Nombre de transcrits exportés suite aux comparaisons de paires de gène avec CG-Alcode. Les résultats concernent les trois espèces d'étude, l'homme (hs), la souris (mm) et le chien (clf). Les chiffres finaux tiennent compte du nombre total de transcrits contenus dans le répertoire de chacune des espèces sans les transcrits redondants.

	Homme	**Souris**	**Chien**
Transcrits initiaux connus (*Ensembl*)	**21**	**13**	**3**
Transcrits exportés dans la comparaison hs-clf	2	-	18
Transcrits exportés dans la comparaison hs-mm	7	11	-
Transcrits exportés dans la comparaison mm-clf	-	1	12
Transcrits redondants	2	1	10
Transcrits prédits	**7**	**11**	**20**
Nombre total de transcrits	**28**	**24**	**23**

3.1.3 CREM : réinjection des prédictions

Après réinjection des prédictions, les modèles GA et GO de chacune des comparaisons sont plus complets grâce aux informations ajoutées avec les transcrits prédits. Pour la comparaison hs-mm, les modèles ont très peu changé suggérant que les modèles contenaient déjà une grande

partie des informations. Ce sont les modèles de la comparaison mm-clf qui ont obtenu davantage de nouvelles informations. La réinjection ajoute donc plus d'informations aux modèles affinant ainsi la structure du modèle du gène.

En conclusion, il est impossible de déduire le modèle commun aux trois espèces de CREM à partir des modèles de paires de gènes, chaque gène à ses spécificités en terme de perte ou de gain. Il faudrait, pour comparer directement les modèles de gènes, définir un algorithme de numérotation multi-gènes des blocs partagés et spécifiques. Nous proposons de représenter l'homologie des trois gènes CREM sous la forme de graphes d'orthologie (figure 12 partie 3.1.4)

Tableau 5 – Modèles GA et GO obtenus par comparaison de paires de gènes pour l'homme, la souris et le chien après réinjection des transcrits prédits. Chaque comparaison de paire de gènes retourne un modèle GA et un modèle GO propre à l'espèce. En rouge figure les blocs spécifiques d'une espèce dans les modèles GO au sein d'une comparaison de paire de gènes pour les modèles GO et en bleu les blocs exoniques obtenus grâce aux transcrits de la deuxième espèce.

homme - souris	
GA homme	[A<.>B[C<.[D<.[E<.>F<.>G<.>H<IJ].>K[L<.[N<.[O<.>Q[R<.>S<.>T].>UV]
GA souris	[A<.>B[C<.[D<.[E<.>F<.>G<.>H<I].>K[L<.[M[N<.[P>Q[R<.>S<.>T].>U]
GO homme	[A<.>B[C<.[D<.[E<.>F<.>G<.>H<IJ].>K[L<.M][N<.[O<.P>Q[R<.>S<.>T].>UV]
GO souris	[A<.>B[C<.[D<.[E<.>F<.>G<.>H<I]J.>K[L<.[M[N<.[P>Q[R<.>S<.>T].>U]V]
homme - chien	
GA homme	[A<.>D[E<.[F<.[G<.>H<.>I<.>J<K].>L[M<.[Q<.[R<.>T[U<.>W<.>X].>YZ]
GA chien	[A<.[B<.>C>D[E<.[F<.[G<.>H<.>I<.>J<K].>L[M<.>N<.>O<.[P[Q<.[S>T[U<.>V>W<.>X].>Y]
GO homme	[A<.B<.>C>D[E<.[F<.[G<.>H<.>I<.>J<K].>L[M<.P[Q<.[R<.[S>T[U<.V>W<.>X].>YZ]
GO chien	[A<.[B<.>C>D[E<.[F<.[G<.>H<.>I<.>J<K].>L[M<.>N<.>O<.[P[Q<.[S>T[U<.>V>W<.>X].>Y]Z]
souris - chien	
GA souris	[A<.>D[E<.[F<.[G<.>H<.>I<.>J<K].>M[N<.[Q[R<.[S>T[U<.>W<.>X].>Y]
GA chien	[A<.[B<.>C>D[E<.[F<.[G<.>H<.>I<.>J<KL].>M[N<.>O<.>P<.[Q[R<.[S>T[U<.>V>W<.>X].>Y]
GO souris	[A<.B<.>C>D[E<.[F<.[G<.>H<.>I<.>J<K]L.>M[N<.[Q[R<.[S>T[U<.>V>W<.>X].>Y]
GO chien	[A<.[B<.>C>D[E<.[F<.[G<.>H<.>I<.>J<KL].>M[N<.>O<.>P<.[Q[R<.[S>T[U<.>V>W<.>X].>Y]

3.1.4 CREM : orthologie des sites fonctionnels

Ces modèles complétés à partir de la réinjection sont utilisés pour la génération d'un graphe de relations d'orthologie entre sites fonctionnels. Le graphe visualisé pour CREM est présenté à la figure 12.

Le graphe révèle 50 sous-graphes dont 60% sont des *clique3*. Ces cas correspondent à des sites fonctionnels partagés entre les trois espèces. D'un point de vue phylogénétique, on dira que ces sites fonctionnels sont présents depuis l'ancêtre *Eutheria*.

Sur ce graphe on retrouve aussi 5 doublons donc 3 qui partagent des informations entre la souris et le chien. Pour ceux-là, le gène humain a donc perdu cette information. Un doublon

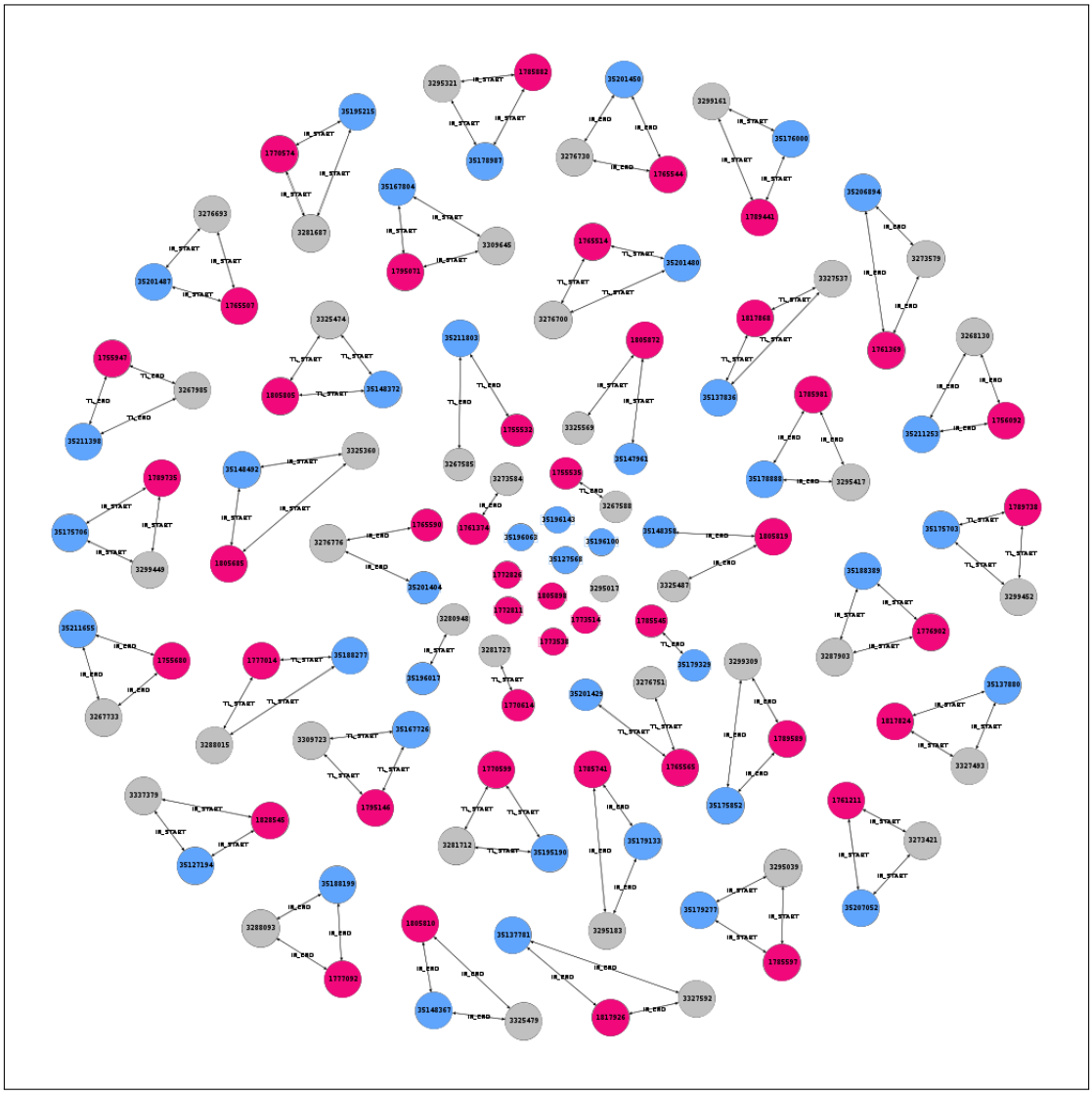


Figure 12 – Graphe des relations d'orthologie entre les sites fonctionnels du gène CREM. Le graphe de relations d'orthologie du gène CREM est constitué d'un ensemble de sous-graphes. Un sous-graphe à trois sommets formant une clique indiquent un site fonctionnel orthologue chez les trois espèces et partagé depuis l'ancêtre *Eutheria*. Les doublons indiquent des pertes et les singletons indiquent des gains de sites (voir partie 2.4.1). Les nœuds sont colorés en bleu, gris et rose correspondant à l'homme, la souris et le chien. La taille des nœuds est proportionnelle au nombre de relations d'orthologie du nœud. Les nœuds représentent les positions génomiques des sites fonctionnels. Un lien entre deux nœuds indique une relation d'orthologie entre les sites fonctionnels de deux espèces.

est retrouvé entre l'homme et le chien, suggérant une perte pour la souris. Quant au dernier doublon, il est retrouvé entre l'homme et la souris, illustrant une apparition de l'information chez l'ancêtre *Euarchontoglires* ou une perte pour le chien.

Ce graphe contient aussi 10 singletons dont 5 pour le chien, 4 pour la souris et 1 pour l'homme. Le chien et la souris ont donc acquis davantage de nouvelles informations, au cours de l'évolution de leur gène, par rapport à l'homme.

Enfin, ce graphe contient aussi 5 cas ambigus (voir partie 2.4.1) révélant une histoire complexe pour le gène CREM.

L'analyse détaillée, pour CREM, est menée à large échelle sur une cohorte de 801 triplets de gènes dont on présente les conclusions principales.

3.2 Etude des 801 triplets de gènes

3.2.1 Exportation des transcrits prédits

Les résultats des transcrits exportés et conservés (déduits des transcrits redondants) dans les répertoires des espèces est détaillé dans le tableau 6. 1 785 transcrits ont été réinjectés dans le transcriptome de l'homme (figure 13), 2 437 ont été réinjectés chez la souris et 2 979 ont été réinjectés chez le chien. Au total, ce sont 7 201 transcrits prédits qui sont réinjectés. Les transcrits exportés complètent les répertoires de transcrits des espèces de 17,53% pour l'homme, de 27,17% pour la souris et de 48,03% pour le chien.

Ainsi, ces réinjections sont un moyen de compléter les transcriptomes des espèces à partir de données déjà connues. Parmi nos trois espèces d'étude, le chien voit son transcriptome presque doubler. La méthode permet ainsi d'annoter les transcriptomes des espèces comme le montre la figure 13 pour le cas de l'homme.

Tableau 6 – Évolution du nombre de transcrits à partir de ceux connus et de ceux exportés suite aux comparaisons de paires de gènes avec CG-Alcode. Les résultats concernent les trois espèces d'étude, l'homme (hs), la souris (mm) et le chien (clf) pour leur 2 167 gènes orthologues. Le chiffre total tient compte des transcrits connus sur *Ensembl*, des transcrits prédits et non redondants.

	Homme	**Souris**	**Chien**
Transcrits initiaux connus (*Ensembl*)	**8 396**	**6 531**	**3 224**
Transcrits exportés dans la comparaison hs-clf	553	-	2 561
Transcrits exportés dans la compraison hs-mm	1 359	2 173	-
Nombre de transcrits exportés dans la comparaison mm-clf	-	516	1 770
Transcrits redondants	127	252	1 352
Transcrits prédits	**1 785**	**2 437**	**2 979**
Nombre total de transcrits	**10 181**	**8 968**	**6 203**

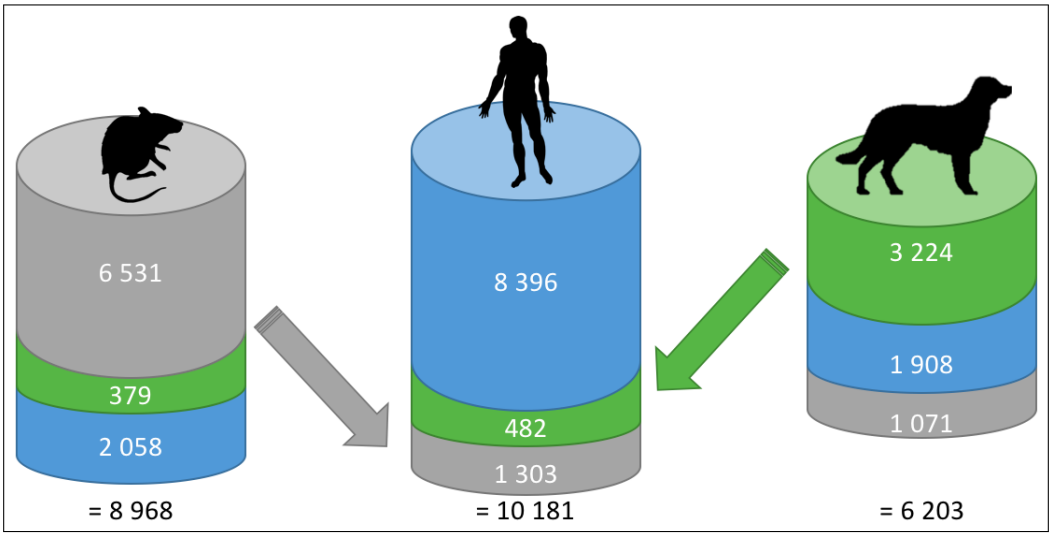


FIGURE 13 – Réinjections des transcrits prédits au travers des comparaisons de paires de gènes. Le premier étage des colonnes correspond aux transcrits connus (*Ensembl*), les deux autres aux transcrits réinjectés à partir des deux autres espèces. Chez l'homme (bleu), 8 396 transcrits sont connus, 482 ont été prédits exprimables à partir des transcrits connus du chien (vert) et 1 303 ont été prédits exprimables à partir des transcrits connus chez la souris (gris). Ainsi, 1 785 transrits prédits sont réinjectés dans le répertoire de transcrits de l'homme qui passa à 10 181 transcrits.

3.2.2 Graphes d'orthologie de sites fonctionnels

Parmi les 2 167 triplets de gènes de départ, 2 148 graphes de sites fonctionnels ont pu être tracés. Les autres graphes qui n'ont pas pu être tracés proviennent du fait que CGALCODE n'arrive pas à aligner des transcrits sur le gène, ce qui engendre une erreur et arrête la comparaison de paires de gènes. Avec cette information manquante, on n'est pas en mesure de tracer de graphe. Parmi les 2 148 graphes, 21 ont été retirés car il s'avère que les données des gènes et des transcrits fournies pour le chien posent problème : on n'obtient pas d'alignements ce qui fait que l'on génère des graphes uniquement entre l'homme et la souris. Enfin, seuls les graphes ne contenant pas de cas ambigus (exemple de CREM, voir partie 3.1.4) ont été conservés pour analyse, réduisant leur nombre à 801 graphes soit 36,96% des triplets de gènes.

L'analyse de la composition de ces 801 graphes est présentée sur la figure 14.

La figure 14a montre la composition générale des 801 graphes. Parmi l'ensemble, 82 graphes (10,23% des graphes) sont composés uniquement de sous-graphes de *clique3*. Ces graphes ont une structure d'ensemble entièrement partagée entre les trois espèces, on dira que leur modèle de structure de gène a donc conservé la même information partagée depuis l'ancêtre *Eutheria*. Ces 82 graphes correspondent à des triplets de gènes orthologues *Eutheria*.

Les autres graphes correspondent à des cas complexes. La figure 14b illustre les 621 graphes de *clique3*-doublons. Parmi eux, 258 graphes sont composés de doublons ne contenant pas de sites fonctionnels chez le chien. Ainsi, les gènes du chien considérés ont des informations qu'ils ont soit perdues soit que l'ancêtre *Euarchontoglires* a acquis.

Le second chiffre important de la figure 14b concerne le nombre de graphes qui ont au moins une information perdue chez la souris (151 et 96). Pour ces gènes, on devrait s'attendre à obtenir une diversité de transcrits murins moins importantes.

Concernant la figure 14c, qui présente les 631 graphes de *clique3*-singletons, on constate que le nombre de graphes n'ayant que des sites fonctionnels uniques chez l'homme ou chez la souris est similaire. On a autant de sites fonctionnels spécifiques à l'homme ou à la souris.

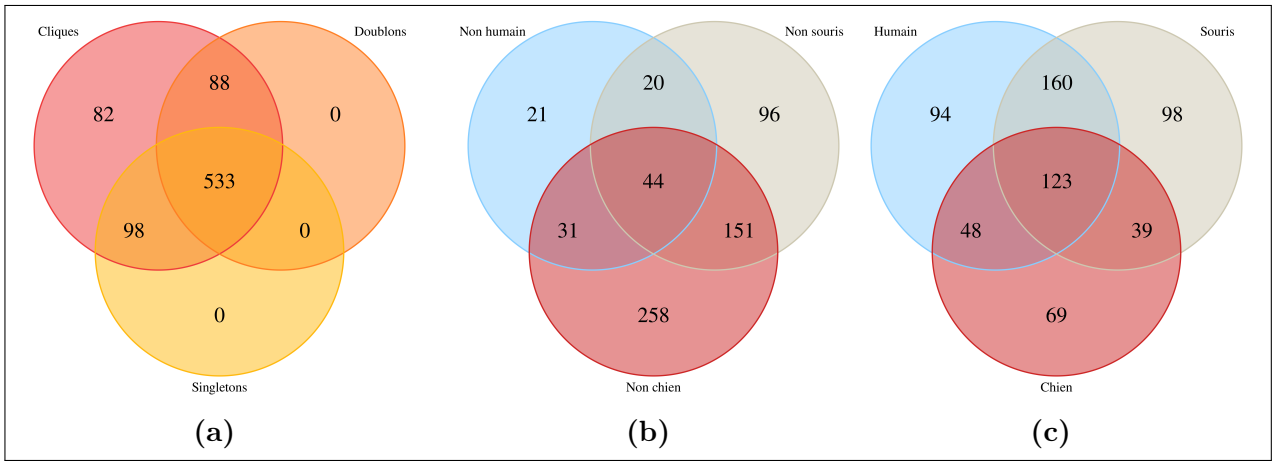


FIGURE 14 – Répartition des 801 graphes de sites fonctionnels. (a) Répartition des composantes entre les 801 graphes conservés. Au total, 621 doublons et 631 singletons sont repartis entre les 801 graphes contenant tous au moins une *clique3*. **(b)** Répartition des doublons dans les 621 graphes entre les espèces. "Non espèce 1" signifie que l'on considère le doublon "espèce 2 - espèce 3" où "espèce 1" n'est pas présente au sein de ce doublon. **(c)** Répartition des singletons dans les 631 graphes entre les espèces. Pour une espèce considérée, on regarde le nombre de fois où le singleton de l'espèce est présent au moins une fois parmi l'ensemble des 801 graphes.

Ainsi, si on regarde la répartition des transcrits orthologues entre les espèces au travers des graphes de relations d'orthologie entre transcrits (figure 15a), on constate qu'un certain nombre de graphes (132) contiennent uniquement des transcrits orthologues partagés entre les trois espèces. Ainsi, le répertoire de transcrits est plus conservé que le répertoire de sites fonctionnels. On a donc des innovations et des pertes récentes pour les autres cas.

De plus, on constate une relation entre le fait que les gènes du chien aient perdu des informations (258 graphes, figure 14b) et le nombre de transcrits qui ne sont pas partagés avec le chien (115 graphes, figure 15b). La perte d'informations dans le modèle d'un gène pour le chien réduit son catalogue de transcrits exprimables par ce gène.

Enfin, on constate dans ces résultats, dans la figure 15c qu'un certain nombre de graphes (41, 39, 43) contiennent uniquement des transcrits spécifiques aux espèces. Or, avec les résultats de la figure 14b, on s'attendait au fait que la grande quantité d'informations perdue par les gènes murins devrait présenter une diversité moins grande des transcrits murins, ce qui n'est pas le cas ici.

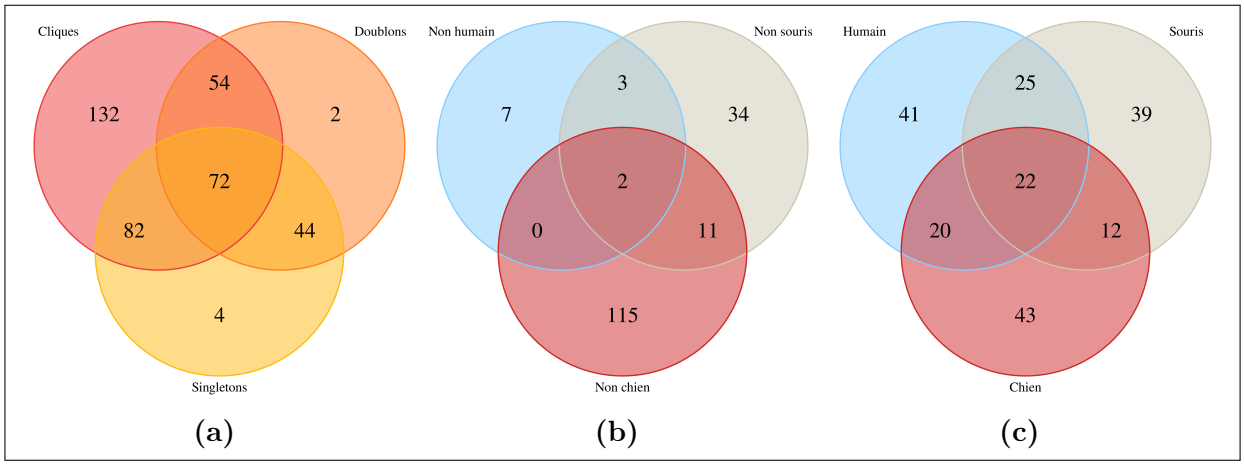


Figure 15 – Répartition des 390 graphes de transcrits. (a) Répartition des composantes entre les 390 graphes conservés. **(b)** Répartition des doublons dans les graphes de transcrits entre les espèces. "Non espèce 1" signifie que l'on considère le doublon "espèce 2 - espèce 3" où "espèce 1" n'est pas présente au sein de ce doublon. **(c)** Répartition des singletons dans les graphes de transcrits entre les espèces. Pour une espèce considérée, on regarde le nombre de fois où le singleton de l'espèce est présent au moins une fois parmi l'ensemble des 390 graphes.

4 Conclusion et Perspectives

CG-Alcode est une méthode qui s'appuie sur les données connues d'*Ensembl*. Ces résultats dépendent ainsi de la quantité d'informations initiale mise à disposition. La prédiction de transcrits par comparaison phylogénétique a ainsi été appliquée à trois espèces d'étude, l'homme, la souris et le chien, sur une cohorte de 2 167 triplets de gènes.

Parmi les trois espèces d'étude, le chien a un répertoire appauvri en comparaison des deux autres espèces. L'idée d'importer les prédictions de la méthode enrichie les transcriptomes des espèces permettant d'annoter de nouvelles informations et de compléter les connaissances disponibles. Au total, 7 201 transcrits ont pu être prédits.

Ainsi, la génération de graphes à partir de ces nouvelles informations permet d'apporter une vision et une interprétation fine sur l'évolution des gènes et sur leurs transcrits exprimés en comparant les trois espèces en même temps. 82 graphes de relations d'orthologie entre sites fonctionnels obtenus ont une structure fine du gène identique entre les trois espèces et qui ont le même potentiel de transcrits exprimables par les gènes.

Pour CREM, 60% de sites fonctionnels sont partagés chez les trois espèces et 10 sites fonctionnels ont des spécificités liées aux espèces.

Néanmoins, les cas de transcription alternative impliquant plusieurs trancrits codants un même ORF devraient à l'avenir être pris en considération dans cette méthode pour augmenter l'impact des résultats et leur précision. En effet, CG-Alcode ne s'appuie pas sur cette information pour effectuer ses prédictions qui concernent à l'heure actuelle seulement les ORF codants et pas les régions non traduites (UTR *Untranslated Transcribed Region*) des ARNm.

A partir de ces graphes, la perspective serait de pouvoir s'y appuyer pour définir un nommage commun des blocs pour des triplets de gènes et même sur un nombre plus important de gènes orthologues et paralogues. Une autre perspective serait de réaliser des analyses phylogénétiques plus poussées pour savoir comment se sont passées les évolutions de perte et de gain d'exons dans les gènes des espèces.

5 Références

[1] Shabalina, S. A., Ogurtsov, A. Y., Spiridonov, N. A. & Koonin, E. V. Evolution at protein ends : major contribution of alternative transcription initiation and termination to the transcriptome and proteome diversity in mammals. *Nucleic Acids Research* **42**, 7132–7144 (2014). URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gku342.

[2] Kelemen, O. *et al.* Function of alternative splicing. *Gene* **514**, 1–30 (2013). URL http://linkinghub.elsevier.com/retrieve/pii/S0378111912009791.

[3] Berget, S. M., Moore, C. & Sharp, P. A. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 3171–3175 (1977). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC431482/.

[4] Chow, L. T., Gelinas, R. E., Broker, T. R. & Roberts, R. J. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* **12**, 1–8 (1977). URL https://www.cell.com/cell/abstract/0092-8674(77)90180-5.

[5] Keren, H., Lev-Maor, G. & Ast, G. Alternative splicing and evolution : diversification, exon definition and function. *Nature Reviews. Genetics* **11**, 345–355 (2010).

[6] Maniatis, T. & Tasic, B. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418**, 236–243 (2002). URL http://www.nature.com/articles/418236a.

[7] Min, F., Wang, S. & Zhang, L. Survey of Programs Used to Detect Alternative Splicing Isoforms from Deep Sequencing Data *In Silico*. *BioMed Research International* **2015**, 1–9 (2015). URL http://www.hindawi.com/journals/bmri/2015/831352/.

[8] Blanchette, M., Green, R. E., Brenner, S. E. & Rio, D. C. Global analysis of positive and negative pre-mRNA splicing regulators in Drosophila. *Genes & Development* **19**, 1306–1314 (2005). URL http://genesdev.cshlp.org/content/19/11/1306.

[9] Khanna, A. & Stamm, S. Regulation of alternative splicing by short non-coding nuclear RNAs. *RNA Biology* **7**, 480–485 (2010). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3062236/.

[10] Buratti, E. & Baralle, F. E. Influence of RNA Secondary Structure on the Pre-mRNA Splicing Process. *Molecular and Cellular Biology* **24**, 10505–10514 (2004). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC533984/.

[11] Tapial, J. *et al.* An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms.

Genome Research **27**, 1759–1768 (2017). URL http://genome.cshlp.org/lookup/doi/10.1101/gr.220962.117.

[12] Ibrahim, E. C. *et al.* Weak definition of *IKBKAP* exon 20 leads to aberrant splicing in familial dysautonomia. *Human Mutation* **28**, 41–53 (2007). URL http://doi.wiley.com/10.1002/humu.20401.

[13] de Miguel, F. J. *et al.* Identification of Alternative Splicing Events Regulated by the Oncogenic Factor SRSF1 in Lung Cancer. *Cancer Research* **74**, 1105–1115 (2014). URL http://cancerres.aacrjournals.org/cgi/doi/10.1158/0008-5472.CAN-13-1481.

[14] Ning, K. & Fermin, D. SAW : A Method to Identify Splicing Events from RNA-Seq Data Based on Splicing Fingerprints. *PLoS ONE* **5** (2010). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2919401/.

[15] Zhou, A. *et al.* Alt Event Finder : a tool for extracting alternative splicing events from RNA-seq data. *BMC Genomics* **13**, S10 (2012). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3535697/.

[16] Sun, X. *et al.* SplicingTypesAnno : Annotating and quantifying alternative splicing events for RNA-Seq data. *Computer Methods and Programs in Biomedicine* **119**, 53–62 (2015). URL http://linkinghub.elsevier.com/retrieve/pii/S0169260715000280.

[17] Liu, Q. *et al.* Detection, annotation and visualization of alternative splicing from RNA-Seq data with SplicingViewer. *Genomics* **99**, 178–182 (2012). URL http://linkinghub.elsevier.com/retrieve/pii/S0888754311002679.

[18] Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods* **10**, 1177–1184 (2013). URL https://www.nature.com/articles/nmeth.2714.

[19] Blanquart, S. *et al.* Assisted transcriptome reconstruction and splicing orthology. *BMC Genomics* **17** (2016). URL http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-3103-6.

[20] Pujar, S. *et al.* Consensus coding sequence (CCDS) database : a standardized set of human and mouse protein-coding regions supported by expert curation. *Nucleic Acids Research* **46**, D221–D228 (2018). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5753299/.

[21] Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Research* **46**, D754–D761 (2018). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5753206/.

[22] Ouangraoua, A., Swenson, K. M. & Bergeron, A. On the Comparison of Sets of Alternative Transcripts. In Hutchison, D. *et al.* (eds.) *Bioinformatics Research and Applications*, vol. 7292, 201–212 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2012). URL http://link.springer.com/10.1007/978-3-642-30191-9_19.

[23] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410 (1990).

[24] Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring Network Structure, Dynamics, and Function using NetworkX 5 (2008).

[25] Bastian, M., Heymann, S. & Jacomy, M. Gephi : An open source software for exploring and manipulating networks (2009). URL `http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154`.

Master de Bioinformatique de l'Université de Rennes 1, année universitaire 2017-2018, Guillaudeux Nicolas

Résumé : Prédiction de transcriptome : analyse comparative multi-gènes, orthologues. La formation des transcrits à partir des gènes eucaryotes est due à un processus d'épissage alternatif qui sélectionne une partie des régions, les exons, dans lesquels certains seront traduits en protéines. Les partenaires de ce stage ont développé une méthode de prédiction des transcrits baptisée CG-Alcode. La méthode CG-Alcode exploite le principe de conservation de séquences d'un même gène entre deux espèces (dits "gènes orthologues"). Ce stage a pour objectif de poursuivre le développement de la méthode afin de l'adapter à de la comparaison multi-espèces en se basant sur trois espèces d'étude (homme, souris et chien). L'étude porte sur une cohorte de 2 167 triplets de gènes orthologues et est illustrée sur un gène en particulier : le gène CREM. Le protocole employé réinjecte les prédictions de transcrits d'un gène source exprimable chez un gène cible dans le répertoire de l'espèce cible. Cela vise à compléter les connaissances actuelles et à annoter le transcriptome des espèces, en particulier celui du chien encore peu documenté. 7 201 transcrits sont ainsi prédits. A partir de là, la construction de graphes de relations d'orthologie a été mise en place pour apporter une vision sur l'évolution des gènes et sur leurs transcrits exprimés. 83 triplets de gènes orthologues ont pu être retrouvés comme présentant une structure fine de gène identique entre les trois espèces, partagée depuis l'ancêtre *Eutheria*. Ces résultats montrent ainsi une première piste de compréhension en terme de gain et de perte d'informations dans les structures de gènes.

Mots-clés : *Orthologie, graphe, CG-Alcode, transcrits, prédiction*

Abstract : Transcriptome prediction : multi-gene comparative analysis, orthologs. The formation of eukaryotic gene transcripts is due to an alternative splicing process that selects some regions, called exons, in which some will be translated into proteins. The partners of this internship have developed a method for predicting transcripts called CG-Alcode. The CG-Alcode method takes advantage of the conservation principle of sequences of the same gene between two species (called orthologous genes). This internship aims to further develop the method in order to adapt it to multi-species comparison based on three study species (human, mouse and dog). The study involves a cohort of 2,167 triplets of orthologous genes and is illustrated on one gene in particular : the CREM gene. The protocol used reinjects transcript predictions of an expressible source gene into the target species' repertory. This aims to complete current knowledge and to annotate the transcriptome of the species, especially the dog transcriptome, which is still little documented. 7,201 transcripts are predicted that way. From there, the construction of orthology relationship graphs was implemented to provide an overview on the evolution of genes and their expressed transcripts. 83 orthologous gene triplets were found as having an identical fine gene structure between the three species, these gene triplets have been shared since the ancestor *Eutheria*. These results thus show a first track of understanding in terms of gain and loss of information in gene structures.

Keywords : *Orthology, graph, CG-Alcode, transcripts, prediction*